МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра математической кибернетики компьютерных наук

РАСПОЗНОВАНИЕ ПАДЕЖЕЙ РУССКОГО ЯЗЫКА ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 – Фундаментальная информатика и информационные технологии
факультета компьютерных наук и информационных технологий
Королькова Максима Олеговича

Научный руководитель:
доцент, к. ф.– м.н. _____ Иванов А.С.
подпись, дата

Заведующий кафедрой:
к.ф – м.н. _____ Миронов С.В.
подпись, дата

Саратов 2019

ВВЕДЕНИЕ

Актуальность темы. Работа посвящена изучению, анализу и поискам решений проблем нейронных сетей. Актуальность, в первую очередь, обусловлена широким спектром возможностей ИНС и повышенным интересом к искусственному интеллекту. ИНС имеют свойства обучаться с помощью анализа входных данных и корректировать ответ в ходе обучения. ИНС, как математическая модель, схожа со строением сетей нервных клеток живых организмов, так как представляет собой систему соединённых между собой процессоров (нейронов), взаимодействующих между собой посредством передачи сигнала. Сигнал, приходящий на вход одному из таких нейронов, специальным образом обрабатывается и передается дальше, на вход следующего нейрона.

Таким образом, такие задачи, как распознавание образов, кластеризация и другие, становятся выполнимыми с использованием ИНС, однако построение таких нейронных сетей – долгий, трудоемкий и сложный процесс.

Цель бакалаврской работы — разработка приложения, распознающего падежи русского языка и предоставляющего статистические данные о ходе эксперимента.

Поставленная цель определила **следующие задачи**:

- рассмотрение особенностей построения и функционирования современных нейронных сетей;
- изучение модели многослойного персептрона;
- создание программного продукта, осуществляющего распознавание падежей русского языка;
- анализ данных и статистик, полученных в ходе работы программы.

Методологические основы «Распознавание падежей русского языка при помощи нейронных сетей» представлены в работах Уоссермен Ф., Хайкин

С., Горбань А.Н., Миркес Е. М., Галушкин А. И., Sebastian Raschka., Гудфеллоу Я., Бенджио И., Курвилль А.

Структура и объём работы. Бакалаврская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 1 приложения. Общий объем работы – 56 страниц, из них 48 страниц – основное содержание, включая 49 рисунков и 1 таблицу, список использованных источников информации – 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Искусственные нейронные сети» посвящен рассмотрению основных понятий. В рамках изучения ИНС было дано определение данного понятия, классификация, применение, обучение ИНС, а также изучен алгоритм обратного распространения.

Искусственная нейронная сеть – набор нейронов, соединенных между собой. Каждый нейрон имеет входы, через которые он принимает сигналы. Поступившие на входы сигналы умножаются на веса связи, по которым они проходят. Все получившиеся произведения поступают в сумматор, на выходе из которого получается взвешенная сумма, после чего сумму подвергают воздействию функции активации.

Классификация ИНС заключается в том, что в ИНС часто используется функция единичного скачка в качестве активационной, которая принимает значение 0 на полуинтервале (-∞; a] и значение 1 на (a; +∞), т.е. меняет своё значение после некоторого заданного порога a, однако чаще применяют сигмоидальную функцию. Существует целый класс сигмоидальных функций. Один из представителей этого класса – логистическая функция

$$out(net) = \frac{1}{1 + \exp(-a * net)}.$$

Логистическая функция является сжимающей, что позволяет вне зависимости от аргумента получать выход в пределах от 0 до 1, более гибкой, чем функция единичного скачка.

Вторым примером сигмоидальной функции, используемой чаще биологами для более реалистичной модели нервной клетки, является гиперболический тангенс.

$$out(net) = \tanh\left(\frac{net}{a}\right).$$

Параметр определяет степень крутизны графика функции. Функция позволяет получать на выходе значения разных знаков, к примеру -1 и 1.

Гиперболический тангенс обладает всеми полезными свойствами логистической функции.

По количеству слоев в нейронной сети выделяют однослойные и многослойные НС. В однослойных нейронных сетях сигналы с входного слоя сразу попадают на выходной слой, который и производит все необходимые вычисления.

Многослойные нейронные сети, помимо входного и выходного слоев, имеют скрытые слои нейронов, их количество варьируется в зависимости от сложности задачи.

По характеру связей в ИНС выделяют сети прямого распространения и с обратными связями. Сети прямого распространения – ИНС, в которых сигнал распространяется строго от входного слоя к выходному.

Сети с обратными связями – сети, в которых сигнал может также распространяться в обратную сторону, от выходов ко входам.

ИНС применяют как инструменты для: распознавания образов и классификации; принятия решений и управления; прогнозирования; сжатия данных.

С ростом требований к ИНС появилась необходимость **обучения сетей,** другими словами изменения весовых коэффициентов так, чтобы рассматриваемая НС становилась наиболее эффективной для решения тех или иных задач.

Существует три разновидности обучения нейросетей – с учителем, без и смешанная.

Теория обучения нейронных сетей выделяет три свойства, тесно связанных с обучением: емкость, сложность образцов и вычислительная сложность.

Также широко известны 4 правила обучения: коррекция по ошибке, машина Больцмана, правило Хебба и обучение методом соревнования.

Самым известным вариантом обучения нейронной сети является – **алгоритм обратного распространения.** В алгоритме вычисляется вектор градиента поверхности ошибок, указывающий направление кратчайшего спуска по поверхности из данной точки, что позволяет уменьшить ошибку.

Алгоритм обратного распространения ошибки состоит из трёх стадий: подача на входы сети обучающих данных, обратное распространение ошибки и корректировка весов.

При этом существуют два варианта выполнения данных стадий в зависимости от подачи обучающей выборки. Первый вариант, когда образцы обучающей выборки подаются по одному на вход ИНС, возможно, в режиме реального времени.

Второй вариант, когда образцы обучающей выборки подаются все сразу, алгоритм высчитывает ошибку на выходе НС для каждого образца, суммирует ошибки и в соответствии уже с этой суммой корректирует весовые коэффициенты.

Второй раздел «Определение падежей с помощью ИНС» включает в себя такие вопросы как, определение падежа с помощью ИНС, сложности при проектировании и примеры ИНС.

Задача определения падежа слова в русском языке – нелегкий и трудоемкий процесс. Традиционный формальный метод, подразумевающий анализ падежных показателей, к примеру флексий, дает систематические сбои в случаях падежной омонимии, которая часто встречается в русском языке.

Падежная система русского языка, применительно к именным частям речи, таким как существительные, прилагательные, числительные, подразумевает разные флексии в зависимости от принадлежности слова к тому или иному склонению и роду, а также его числа (единственное или множественное).

Таким образом, задача определения падежа слова русского языка усложняется совпадением окончаний слов у разных падежей, а как известно, при частом совпадении признаков у разных классов ИНС будет соотносить некоторые образцы с несколькими классами.

Определение склонения существительного в русском языке сводится к нахождению именительной формы единственного числа данного слова и дальнейшему определению окончания в этой форме, и обуславливается это правилом для склонений языка.

Главными задачами при проектировании ИНС для определения падежей являются задание правильных входных данных, выбор правильных параметров сети и поиск идеального решения задачи.

Выбор правильных параметров сети приводит к задаче нахождения баланса между способностью сети выдавать верный ответ на входные данные, используемые в обучающей выборке, и схожие, но неидентичные им данные, принадлежащие тестовой выборке.

Также были рассмотрены **примеры ИНС** распознавания падежей русского языка от сторонних разработчиков. В дипломной работе были проанализированы такие ИНС как: MyStem от компании Yandex и rnnmorph, выполненный в виде библиотеки для языка python.

Анализ показал, что каждое рассмотренное программное решение является недостаточно точным, также существуют более успешные продукты, обычно не предоставляемые для открытого доступа. Однако ни один из продуктов не имеет ни удобного интерфейса, ни статистических данных своей работы.

Третий раздел «Практическая часть» посвящен созданию программы, реализующей нейронную сеть с возможностью изменения параметров, распознающей падежи русского языка и предоставляющей статистические данные о ходе эксперимента.

Используемое программное обеспечение - Windows 10 64-bit.

Используемый язык программирования – Python 3.7.

Для работы программы необходимо в дистрибутиве с продуктом создать папку «Resourses».

В папке «Resourses» хранятся такие параметры, как фактор, функция активации, количества скрытых нейронов и слоев, под соответствующими файлами factor.txt, func.txt, hiddenNeur.txt и hidrow.txt. При запуске программы происходит проверка данных в этих файлах на соответствие.

В дипломной работе был представлен подробный алгоритм использования программы.

Во вкладке «Настройки», следует выбор функции активации из двух предложенных, бинарный и биполярный сигмоиды, реализованный в виде кнопок.

Далее следует окно ввода ограничения по коэффициенту ошибки.

На вкладке «Обучение» можно задать слово для статистики, для которого будет составлен график изменения величины ошибок в ходе обучения сети.

При нажатии кнопки «Запуск обучения» программа выводит окно выбора файла с обучающей выборкой.

Также существуют требования к файлу обучающей выборки. Файл должен являться текстовым, каждое слово записано в отдельную строчку вместе с числовыми соответствиями (1 или 0, через пробел) падежам в порядке: Именительный, Родительный, Дательный, Винительный, Творительный и Предложный, где 1 – значит слово указанного падежа, 0 – обратное.

При успешном выборе файла и нажатии кнопки «Открыть» НС начнет обучение с заданными параметрами, сгенерировав новые веса в случае первого запуска. По истечении обучения появится окно оповещения об окончании с указанием количества пройденных циклов.

Также после обучения в текстовой области появятся записи о том, что НС была обучена на словах, присутствующих в обучающей выборке

Далее, на вкладке – «Ввод данных», требуется ввести слово, для которого необходимо проверить значения уверенности в принадлежности слова к падежам, и нажать кнопку «Ввод».

Следующим шагом необходимо рассмотреть некоторые интересные случаи работы НС Падежи, проанализировать полученные данные и сделать выводы.

Было проведено 5 экспериментов и выявлен 1 наиболее удобный и оптимальный вариант параметров.

Первый эксперимент – стандартные параметры.

Параметры:

- количество нейронов на скрытый слой – 7;
- количество скрытых слоев – 1;
- размерность фактора – 0.1;
- активационная функция – бинарный сигмоид;
- ограничение по коэффициенту ошибки – 0.05;
- количество циклов – 14000.

Увеличение количества нейронов на скрытый слой повлияло, в первую очередь, на требуемое количество циклов для обучения НС с учетом ограничения коэффициента ошибки.

Замечено сильное увеличение времени, затрачиваемого на обучение НС, однако точность определения падежей однозначно улучшилась. С ростом количества нейронов на скрытый слой требуется намного более долгое изменение весов на каждой связи нейронов.

Второй эксперимент – изменение количества нейронов на скрытый слой.

Параметры:

- количество нейронов на скрытый слой – 8;
- количество скрытых слоев – 1;
- размерность фактора – 0.1;
- активационная функция – бинарный сигмоид;
- ограничение по коэффициенту ошибки – 0.05;
- количество циклов – 14000.

НС Падежи, отработав 9042 цикла, обучилась с ограничением коэффициента ошибки 0.05. Степень уверенности данной НС отличается от представленной ранее.

При увеличении количества скрытых слоев увеличивается точность определения падежей, однако сильно растет время работы и необходимое количество циклов обучения. Таким образом, определено оптимальное количество скрытых слоев – 2.

Третий эксперимент – изменение количества скрытых слоев.

Параметры:

- количество нейронов на скрытый слой – 6;
- количество скрытых слоев – 2;
- размерность фактора – 0.1;
- активационная функция – бинарный сигмоид;
- ограничение по коэффициенту ошибки – 0.05;
- количество циклов – 14000.

В среднем обучение проходило за 7673 циклов, однако по времени длилось, разумеется, дольше в полтора раза, так как связей стало во столько же раз больше. При продолжении обучения до четырнадцати тысяч циклов ответы НС сильно приблизились к заданным на обучении

Четвертый эксперимент – изменение фактора.

Параметры:

- количество нейронов на скрытый слой – 6;
- количество скрытых слоев – 2;
- размерность фактора – 0.5;
- активационная функция – бинарный сигмоид;
- ограничение по коэффициенту ошибки – 0.05;
- количество циклов – 14000.

При факторе 0.5 обучение НС происходит за 1435 циклов

Пятый эксперимент – изменение активационной функции.

Параметры:

- количество нейронов на скрытый слой – 6;
- количество скрытых слоев – 2;
- размерность фактора – 0.5;
- активационная функция – биполярный сигмоид;
- ограничение по коэффициенту ошибки – 0.05;
- количество циклов – 14000.

НС с биполярным сигмоидом быстро обучалась, в среднем за 2400 циклов. После 14000 циклов обучения выдает лучшие результаты в сравнении с предыдущими экспериментами.

Таким образом, наиболее удобным и оптимальным вариантом являются параметры:

- количество нейронов на скрытый слой – 6;
- количество скрытых слоев – 2 - 4;
- размерность фактора – 0.5;
- активационная функция – биполярный сигмоид;
- ограничение по коэффициенту ошибки – 0.05.

После получения наиболее оптимальных параметров, был рассмотрен пример работы программного продукта на больших обучающих выборках.

Сложности определения падежей слов русского языка описаны во втором разделе данной работы. Учитывая их, создание НС, точно и

однозначно определяющей падеж, это нелегкая, требующая знаний в лингвистике задача.

В результате работы мы получаем программный продукт, реализующий НС, определяющую падеж слов русского языка, а также позволяющий проводить исследования с НС и находить лучшие параметры для определенных целей.

ЗАКЛЮЧЕНИЕ

Распознавание падежей важный и трудоемкий процесс, и до сих пор не создан универсальный, открытый и достаточно точный продукт для решения данной задачи.

В ходе работы были рассмотрены особенности искусственных нейронных сетей, возможности реализации и проблемы, возникающие в ходе исследования и обучения.

Кроме того, были проанализированы программные решения сторонних разработчиков, осуществляющие распознавание падежей русского языка. В ходе исследований было установлено, что данные продукты не обеспечивают точного определения падежей, а также в большинстве случаев являются закрытыми коммерческими проектами.

В результате был создан программный продукт, реализующий искусственную нейронную сеть с заданными параметрами, которая распознает падежи русского языка, приложение также предоставляет статистические данные о ходе эксперимента.

Основные источники информации:

1. Уоссермен Ф. Нейрокомпьютерная техника: Теория и практика / Пер. на рус. Яз., Зуев Ю.А., Точенов В.А., 1992. 184с.

2. Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей / Сибирский журнал вычислительной математики, 1998, т. 1, № 1. С. 12—24.

3. Хайкин С. Нейронные сети: полный курс / Пер. с англ. Куссуль Н.Н., Шелестова А.Ю., 2006. 1104с.

4. Миркес Е. М., Нейрокомпьютер. Проект стандарта / Новосибирск: Наука, Сибирская издательская фирма РАН, 1999. 337с.

5. Галушкин А. И. Синтез многослойных систем распознавания образов / М.: Энергия, 1974. 368с.

6. Sebastian Raschka. Python Machine Learning / Packt publishing. Birmingham, 2015. P. 18

7. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / М.: ДМК-Пресс, 2017. 652с.